Differences between Swiss High German and German High German via data-driven methods

Schneider, Gerold

Abstract: This study uses data-driven methods to detect and interpret differences between the High German used as standard language of written communication in Switzerland, and German High German. The comparison is based on a comparable web corpus of two million sentences, one million from Switzerland and one million from Germany. We describe differences at the levels of lexis, morphosyntax, and syntax, and compare to previously described differences. We show that data-driven methods manage to detect a wide range of differences.

Differences between Swiss High German and German High German via data-driven methods

Gerold Schneider
Institute of Computational Linguistics
and English Department
University of Zurich

gschneid@ifi.uzh.ch

Abstract

This study uses data-driven methods to detect and interpret differences between the High German used as standard language of written communication in Switzerland, and German High German. The comparison is based on a comparable web corpus of two million sentences, one million from Switzerland and one million from Germany. We describe differences at the levels of lexis, morphosyntax, and syntax, and compare to previously described differences. We show that data-driven methods manage to detect a wide range of differences.

1 Introduction

While the various spoken dialects of Swiss German differ considerably from High German, and also from each other, the differences between Swiss High German and German High German are relatively small, and mostly concern the level of lexis and morphosyntax. Many of the differences have been described in lexica, see e.g. Meyer (1989). Also some syntactic differences have been reported (Dürscheid et al., 2015). There are virtually no structures that are used exclusively in one of the two compared varieties, but preferences for certain constructions and lexical items exist. As the differences are often subtle and small and existing resources are incomplete, a data-driven approach using a large amount of carefully compiled data is recommendable for their detection, and hitherto missing for Swiss German. Our research question is whether data-driven methods are able to find linguistically meaningful differences.

In: Mark Cieliebak, Don Tuggener and Fernando Benites (eds.): Proceedings of the 3rd Swiss Text Analytics Conference (SwissText 2018), Winterthur, Switzerland, June 2018

2 Data and Methods

We apply a data-driven method to the detection of differences. In data-based approaches, existing hypothesis are tested, whereas in data-driven approaches, hypotheses arise from the data. Data-driven methods have the advantage that previously unnoticed differences may be detected, thus improving the recall of the phenomena under observation, potentially showing new patterns that one is not yet aware of, and it also allows one to put the differences into a quantitative perspective. Data-driven methods also have disadvantages, in particular that they depend directly on the corpus and its sampling (Tognini-Bonelli, 2001), that quantitatively rare differences are hard to detect, and that subtle differences may not leave traces on the surface and thus remain unnoticed. In order to partly alleviate the latter, we use morphologically and syntactically annotated data.

2.1 Data: the Wortschatz Leipzig Corpus

The Wortschatz Leipzig corpus[1] consists of a collection of news and web-derived corpora, each comprising one million sentences. For the comparison of Swiss High German to German High German, we use their Swiss and matching German web corpus, which contain random texts from the year 2002. The Swiss corpus contains 15.817.004 words, the German one 16.850.144 words.

2.2 Lemmatisation, Tagging and Parsing

As pre-processing steps, we reduced full-forms to their lemmas and applied part-of-speech tagging. For these two steps we used Treetagger (Schmid, 1994),

[1] http://wortschatz.uni-leipzig.de/de

which employs the STTS tagset (Schiller et al., 1995). To allow comparability, we have also mapped all occurrences of *ß* to *ss*, as Swiss High German does not use *ß*.

We use a syntactic dependency parser (Sennrich et al., 2009) for the step of syntactic annotation. The set of syntactic dependency labels is described in Foth (2006). Although automatic annotation is not error-free, the levels of noise can now be considered low enough to profit from these resources (see e.g. van Noord and Bouma (2009)).

2.3 The method of document classification

Document classification is a supervised method generally used to assign each document, whether a newspaper article, a web page, a book, a paragraph, a tweet, or a similar discourse unit, to a class. Classes can, for example, be broad topics divided into the binary classes of relevant or irrelevant documents for an Information Retrieval task (see Jurafsky and Martin (2009, chapter 23.1)) or Manning and Schütze (2001, chapter 25)) for an introduction). In the majority of the implementations, the words in the documents are used as discriminators between the classes, typically without respecting their sequence or syntactic context, which is why the method is called a "bag-of-words" approach. Since every word type (as soon as it reaches a token frequency above a certain threshold) is a feature, there are often thousands of features. Each feature in isolation is usually neither a good descriptive feature nor a good discriminator between the classes. The simplest approach, Naïve Bayes, simply gives equal weight to each feature. More advanced algorithms, for example logistic regression, which we use in the present study, give optimal weight to each feature. While most features in isolation are bad discriminators, some are better and logistic regression automatically finds the optimal weight (also called influence) for each feature. Those features that obtain a high weight are relatively good discriminators and therefore they can be considered typical of their class. These words can be interpreted as keywords, because document classification is also a possible keyword extraction algorithm (Yang et al., 2013).

For the detection of English National Dialects, Lui and Cook (2013) have tested a range of methods. They conclude that document classification performs best on the task of detecting the originating nation of a text, they state that this is probably because of the very large feature set of this method. The main interest of Lui and Cook (2013) was to obtain a high classfication accuracy, they were not mainly interested in lingustic interprations of the features.

2.4 Overuse metrics

As overuse metric we use O/E and derived measures. O/E stands for Observed divided by Expected, where Expected is the homogenous distribution over the entire corpus. The value gives a direct, and easily interpretable effect size. O/E is often affected by sparse data problems which can lead to inflated values for items with low counts. One thus sometimes gets a more accurate impresssion by adding a frequency factor, for example O^2/E or $O * log(O)/E$. O/E is well known from research on collocations (Evert, 2009), but it is a useful general overuse metric. The ranking of features which O/E delivers is identical to the one obtained by Mutual Information (MI), a popular metric in Information Theory (Shannon, 1951; Cover and Thomas, 1991), but even easier to interpret.

3 Results

3.1 Lexis

In a first classification task, we applied document classification to the raw texts. Our tool of choice for applying the method is LightSide[2] because it is easy to use, includes tokenisation, and offers a wide range of machine learning algorithms, including logistic regression from the LIBLINEAR library (Fan et al., 2008). It also performs cross-validation automatically. We used 5-fold cross validation. We formed pseudo-documents of 100 random sentences each, delivering 10000 Swiss High German and 10000 German High German documents, and have set the minimum frequency for words to 50, which delivers over 20000 bag-of-words features.

While the performance of the system is near-perfect (only one document was misclassified), the strongest features are dominated by place names and proper names. Therefore, as a second classfication task, we thus removed all proper names (tag *NE*), and we also replaced full forms by lemmas. The classification is still very accurate (99.97% accuracy, 8 documents are misclassified), and a large subset of the top

[2] http://ankara.lti.cs.cmu.edu/side/

Position	Feature	Frequency (CH)	Feature Influence ↓	Comment
6	welch	9664	11.595	Relative Pronoun
7	zürcher	1869	11.161	züricher
14	basler	1324	8.798	baseler
15	galle	1366	8.727	(dialect word not recognized as proper name)
16	gemäss	2424	8.624	zufolge
17	anlass	2660	8.572	veranstaltung
18	lehrperson	1052	8.259	
19	gemeinderat	1668	8.123	
20	allfällig	982	8.106	etwaig
22	selber	3159	7.855	selbst
25	innert	1058	7.498	binnen, innerhalb
26	generalversammlung	995	7.254	jahreshauptversammlung
28	präsident	2269	7.139	vorsitzende/r
29	spital	923	7.107	krankenhaus
30	zudem	3463	7.066	ansonsten
31	stadtrat	1272	6.929	
32	dank	4084	6.801	aufgrund
34	via	1410	6.778	
35	nebst	1146	6.686	neben; ausserdem
36	eidgenössisch	937	6.658	
37	divers	1930	6.239	unterschiedlich
38	benützen	673	6.123	anwenden
40	person	5292	6.093	
41	kurs	3043	6.058	lehrgang
42	verschieden	7429	6.028	anders
43	bedürfnis	2386	5.949	erfordernis
44	gratis	985	5.928	kostenlos
45	art.	1065	5.886	
46	schulhaus	726	5.876	gesamtschule
47	resp	690	5.864	
49	ferien	900	5.731	urlaub
50	rasch	1584	5.669	schnell
51	gemeinde	4067	5.609	
52	schülerin	1717	5.502	
53	vgt	645	5.305	
54	bezüglich	1189	5.294	hinsichtlich
55	ur	678	5.262	
56	vermehrt	894	5.204	verstärkt, gehäuft
57	anliegen	1241	5.077	
58	pro	4271	4.989	zum vorteil von
59	verlangen	2403	4.956	begehren
60	besuchen	3047	4.873	
61	junior	752	4.866	
62	falls	2592	4.835	
63	definitiv	822	4.789	abschliessend, bestimmt
65	stiftung	1239	4.699	
66	laufend	1961	4.665	
68	velo	495	4.644	fahrrad
69	statut	672	4.642	satzung
70	tier	3806	4.641	
71	bundesamt	672	4.628	
72	publizieren	731	4.628	veröffentlichen

Table 1: Top-weighted 72 features indicating Swiss High German

f(CH)	POS tag	f(DE)	$O/E \downarrow$	O^2/E	$O * logO/E$	Comment
3191	PWAT	2387	1.144	3651	4.009	Relative pronoun *welche/r/s*
40890	TRUNC	39429	1.018	41634	4.696	
67807	VAINF	66674	1.008	68378	4.872	Present perfect
292929	APPRART	289036	1.007	294889	5.503	Contraction
19900	PWAV	19662	1.006	20020	4.325	*wo, wobei*
347021	VVPP	343283	1.005	348900	5.570	Present perfect
27732	VVIZU	27642	1.002	27777	4.450	Hedging phrases, Swiss indirectness?
9433	VMINF	9459	0.999	9420	3.969	Present perfect
978574	ADJA	994622	0.992	970615	5.942	
478239	ADJD	487708	0.990	473551	5.624	
1678008	ART	1719480	0.988	1657525	6.149	det+proper name
584421	KON	599329	0.987	577061	5.694	paratactic style
531017	VAFIN	546238	0.986	523514	5.644	Present perfect
1315882	APPR	1366228	0.981	1291182	6.004	synthetic, genitive drop, fewer postpositions
91608	PTKZU	95449	0.979	89727	4.860	Hedging phrases, Swiss indrectness?
92470	PDAT	96374	0.979	90558	4.863	
91277	KOKOM	96083	0.974	88936	4.833	
3752973	NN	3984207	0.970	3640811	6.378	
1123935	$.	1198501	0.968	1087849	5.856	shorter sentences
...						
1801	APPO	2175	0.906	1632	2.949	more postpositions in German High German
287794	CARD	347751	0.906	260644	4.944	
95333	PIS	116763	0.899	85701	4.476	
467198	NE	586103	0.887	414457	5.029	Acronyms such as KFZ
2651	PRELAT	3536	0.857	2272	2.934	fewer *wo, wobei*, fewer relatives with *welcher*
30	PPOSS	41	0.845	25	1.248	
7510	VVIMP	10578	0.830	6236	3.218	German directness? / short forms
2474	PTKANT	3850	0.782	1936	2.655	
1896	ITJ	3035	0.769	1458	2.521	
35	VMPP	59	0.745	26	1.150	absent in CH, rare in DE
545	VAIMP	1262	0.603	329	1.651	German directness? / short forms

Table 2: Overused POS-tags in Swiss High German (and German High German, see bottom), sorted by O / E

features are linguistically meaningful, exhibiting helvetisms. Table 1 lists the top 72 features (of a total of 21789). We have manually filtered non-linguistic features (such as the adjective *Schweizer*) in this list[3], but give the position in the original list in the first column, to give an impression of the level of noise. The last column gives our interpretation, explanation, or near-synonyms which are overused in the German High German corpus, and at least partly explain the high weight of the feature in the Swiss German corpus.

The top entry is the lemma *welch*, stemming from the full forms *welche, welcher, welches* which are strongly overrepresented in Swiss High German, while the relative pronoun forms *der, die, das* are used less often than in German High German. The preposition *gemäss* is overrepresented in Switzerland, while the semantically largely corresponding pre- or postposition *zufolge* is a strong feature of German High German. The majority of the features can be explained, but data-driven approaches also lead to some results which are difficult to explain. Words related to education and schooling (for example *Schülerin*) seem generally overrepresented in the Swiss data. This may be due to a bias in the corpus collection or due to the importance that Switzerland gives to education.

Further down in the list than shown in Table 1, we can still find many lexical differences, for example the Swiss *Velo* for *Fahrrad*, *Offerte* for *Angebot*, *benützen* for *anwenden*, *Gesuch* for *Antrag*, *selber* for *selbst*, *Reservation* for *Reservierung*, *Mitgliederbeitrag* for *Mitgliedsbeitrag*, *Ferien* for *Urlaub*, and hundreds more.

Most of the reported lexemes are not exclusive to one variety, but exploring the feature weights offers an exciting resource to the lexicographer. We can equally browse the strongest German High German features and learn for example that German High German (over-)uses *Personalausweis* for Swiss *Identitätskarte*, *gezahlt* for *bezahlt*, *zeitnah* for *bald*, *Stadtmitte* for *Zentrum*, *PKW* for *Auto*, *Festplatte* for *Harddisk*, *Rundfunk* for *Radio*, *Renovierung* for *Renovation*, etc. Many of these differences are known, but a data-driven approach allows us to verify lists and complete dictionaries.

[3] *zürcher* and *basler* are kept, because they are linguistically meaningful, German High German would use *züricher* and *baseler*

3.2 Morphosyntax

Some morphosyntactic differences are also well known. For example, due to the fact that the Swiss dialects do not use the simple past tense, overuse of the present perfect can be expected. To obtain a more complete picture, we have sorted all part-of-speech tags by overuse metrics. The results sorted by O/E are given in Table 2. The expected overuse of present perfect is mirrored by more auxiliary verbs and participles (*VAINF, VAFIN, VVPP*). The table reflects relative pronouns with *welch* again (*PWAT*), and with *wo* (*PWAV*). The increased frequency of *PWAV* is partly also due to *wobei*, which is described as a Swiss feature in Dürscheid et al. (2015, 228). Determiners (*ART*) are more frequent in Swiss texts because proper names are often preceded by determiners. The fact that infinitive verbs with particle *zu* (*PTKZU, VVIZU*) are more frequent in the Swiss corpus is due to the frequent use of fixed semi-modal, hedging phrases:

(1) Es_PPER ist_VAFIN anzunehmen_VVIZU ,_$, dass_KOUS das_ART Quartier_NN etappenweise_ADJD überbaut_VVPP wird_VAFIN ._$.

(2) Zunächst_ADV ist_VAFIN festzuhalten_VVIZU ,_$, dass_KOUS der_PDS vermeintlich_ADJD ”_$(normale_ADJA ”_$(Zustand_NN nicht_PTKNEG existiert_VVPP ._$.

(3) Der_ART Name_NN ist_VAFIN zurückzuführen_VVIZU auf_APPR ein_ART Treffen_NN von_APPR Veteranen_NN der_ART American_NN Legion_NN 1976_CARD in_APPR den_ART USA_NE ._$.

Semi-modal patterns of the type *es ist _VVIZU* are 50% more frequent in the Swiss than in the German corpus.

It is tempting to interpret these phrases in contrast to the fact that imperatives (*VAIMP, VMIMP, VVIMP*) are among the strongest German High German features. Inspecting the data reveals, however, that the majority of verbs tagged as imperative are in fact short forms, not imperatives, such as

(4) hab_VAIMP ich_PPER einen_ART von_APPR der_ART Sicherheit_NN getroffen_VVPP ._$.

Data-driven approaches are relatively susceptible to skews and systematic errors in the data and the annotation process.

The Swiss texts have shorter sentences, therefore the full stop tag *$.* is also overrepresented in the Swiss data. As Swiss sentences are shorter, the Swiss corpus is also a bit smaller, which explains why most O/E values are slightly below 1.

f(CH)	Prep	f(DE)	O/E	O^2/E	$O * logO/E \downarrow$	comment
1174	nebst	119	1.935	2132	5.574	neben
3421	dank	1007	1.646	5286	5.461	aufgrund
2500	gemäss	688	1.671	3921	5.329	zufolge
158091	mit	152419	1.085	160979	5.294	
232932	in	246247	1.036	226459	5.218	
117727	im	111524	1.094	120912	5.208	
122237	für	121910	1.067	122401	5.094	
135391	von	140765	1.045	132756	5.032	genitive drop
396	ausschliesslich	14	2.058	765	5.018	
1537	via	434	1.661	2397	4.970	
107046	auf	114011	1.032	103673	4.871	
58357	an	59332	1.056	57874	4.727	dative shift
49882	nach	49588	1.068	50029	4.712	
5505	während	3281	1.335	6898	4.688	
38989	am	37637	1.084	39677	4.672	progressive form
59418	bei	63184	1.033	57593	4.627	
22013	vom	19528	1.129	23330	4.602	genitive drop
5464	pro	3620	1.282	6573	4.496	
41888	über	45144	1.025	40321	4.449	
40445	zum	43382	1.028	39028	4.445	
1232	bezüglich	487	1.527	1766	4.430	
37494	zur	40426	1.025	36083	4.402	
24318	vor	25228	1.046	23871	4.305	
14398	beim	13443	1.102	14892	4.301	
19112	unter	19305	1.060	19016	4.260	
1970	mittels	1112	1.362	2518	4.212	
...						
19	fern	38	0.667	13	0.853	
20	kraft	45	0.615	12	0.801	
21	i.	55	0.553	12	0.731	
39	gen	136	0.446	17	0.709	
52	binnen	201	0.411	21	0.705	
15	hinterm	39	0.556	8	0.653	
28	unterm	109	0.409	11	0.592	
6	rechts	11	0.706	4	0.549	
6	links	15	0.571	3	0.445	
6	nah	16	0.545	3	0.424	
28	o.	174	0.277	8	0.401	
13	vorm	95	0.241	3	0.268	
4	vorbehaltlich	25	0.276	1	0.166	
2	übern	12	0.286	1	0.086	
1	überm	10	0.182	0	0.000	
1	unbeschadet	15	0.125	0	0.000	

Table 3: Overused prepositions in Swiss High German (and German High German at the bottom), sorted by O*log(O) / E)

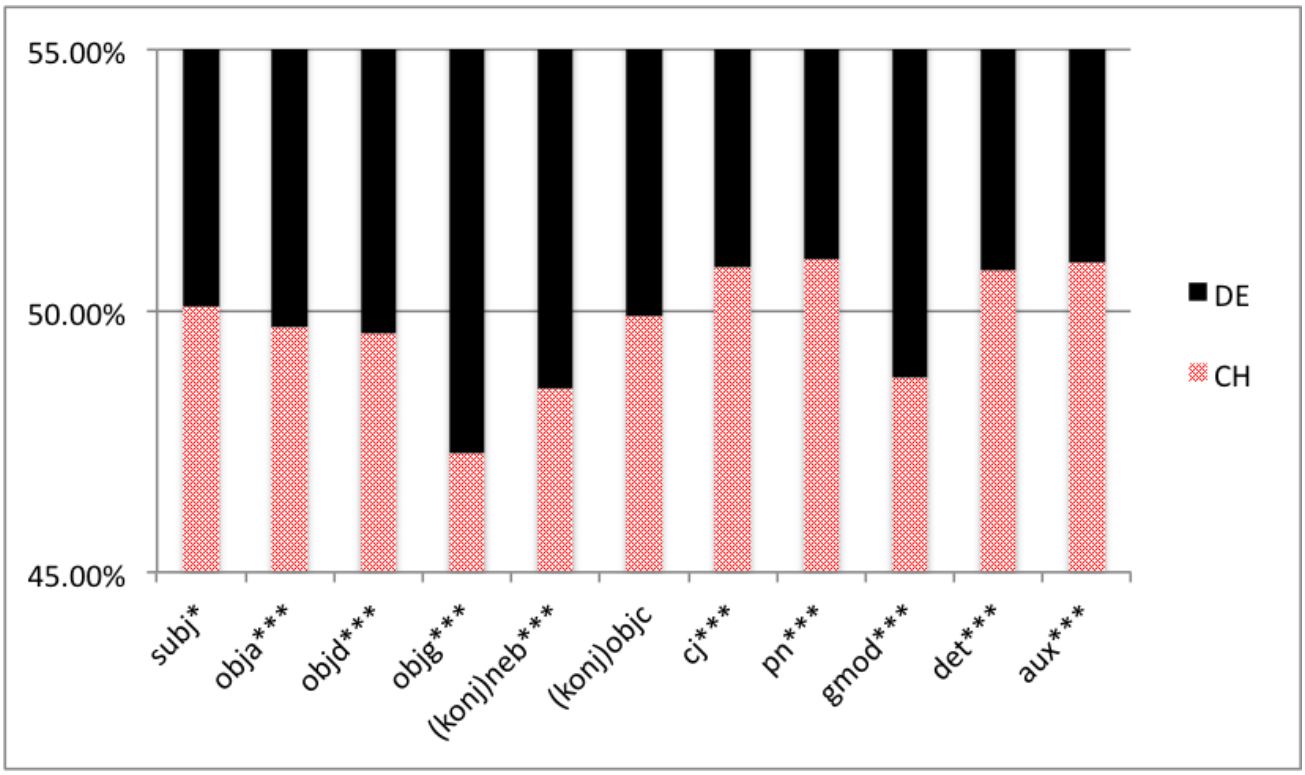

Figure 1: Relative frequencies of important dependency relations. Significance codes of Chi-square contingency test with Yates' continuity correction: *:$p < 5\%$, **: $p < 1\%$, ***: $p < 0.1\%$. The dependency labels are: subj=Subject, obja=Accusative Object, objd=Dative Object, objg=Genitive Object, (konj)neb=Nebensatz/Adjunct Clause, (konj)objc=Complement Clause, cj=Conjunction, pn=Preposition, gmod=Genitive Modification, det=Determiner, aux=Auxiliary Verb)

Prepositions seem to show important dfferences. Contraction of prepositions plus article (*am, im, beim* etc.) are more frequent in the Swiss variety, while postpositions (*APPO*) are a German High German feature. We have already seen in the comparison of lexis that the preposition *gemäss* has the semantically largely corresponding pre- or postposition *zufolge* in German High German, there may be an interdependence. As prepositions (*APPR*) appear among the overused words, we list results of the uses of individual prepositions in Table 3. Some contractions are more typical for Swiss (*im, am, vom, zum, zur, beim*), while others are typically German (*hinterm, unterm, vorm, übern, überm*).

The higher frequency of *am* can also be explained by the frequent use of the progressive form *am*, e.g. *Ich bin am Laufen* (van Pottelberge, 2004), which is more frequent in Swiss High German (Rimensberger, 2014, 107).

As Swiss German has no genitive form, genitive drop is probably also frequent in Swiss High German, explaining the overuse of *von* and *vom*. But in order to investigate this question, syntactic information is needed, which we provide in the following subsection.

3.3 Syntax and Style

The overuse of the tags *VVIZU* and *PTKZU* in Swiss High German, and the frequent conjunctions (*KON*) are stylistic features. Also many of the syntactic differences can be seen as differences in style. On the surface, it is already noticeable that German High German sentences are considerably longer (mean of 15.85 words per sentence) than Swiss High German ones (mean of 14.82 words). Sentence length and complexity are usually strongly correlated. One feature of sentence complexity is the use of subordinating and coordinating clauses. On the level of tags, we can see in Table 2 that conjunctions (*KON*) are even overused in Swiss High German, which is surprising. Looking at the parsed data can give more de-

Variety	wegen des	wegen dem
DE	303	32
CH	213	110

Table 4: Frequencies of *wegen dem* and *wegen des* in the Swiss and the German Corpus (Differences highly significant, $p < 0.01\%$, Chi-square contingency test with Yates' continuity correction).

tailed answers. A comparison of important syntactic relations is shown in Figure 1. Subordinate clauses are expressed by the dependency relations $(konj)neb$ for adjunct clauses and $(konj)objc$ for complement clauses. Particularly adjunct clauses are indeed underused in the Swiss texts. There seems to be a slight trend towards a more paratactic style at the expense of hypotactic style.

Due to the fact that Swiss High German uses more present perfect forms instead of the simple past, the higher frequency of aux dependencies is expected.

There is an overuse of det, which is due to proper names with determiners, as seen in overused POS tags.

Subjects ($subj$) are distributed homogeneously. Accusative objects ($objа$) and dative objects ($objd$) are almost as frequent in Swiss High German as in German High German, but genitive objects ($objg$) are considerably rarer. Prepositions and verbs governing a genitive object, as in the following German High German examples, are rarer in the Swiss texts.

(5) *Trotz_APPR des_ART Zugewinns_NN* ging_VVFIN Karl_NE Braun_NE leer_ADJD aus_PTKVZ ._$.

(6) Da_ADV machte_VVFIN sich_PRF auch_ADV ans_APPRART Werk_NN das_ART Ratsgymnasium_NN ,_$, denn_KON es_PPER *gedachte_VVFIN des_ART alten_ADJA Schlagers_NN* "_$(Für_APPR Gaby_NE tu_VVFIN ich_PPER alles_PIS "_$(._$.

Verbs governing genitive objects are also receding in Standard German (see e.g. Ueberwasser (2014); Schätzle (2013)), but prepositions governing genitive case remain stable across time. Nouns can be modified by genitive NPs (dependency label $gmod$), a phenomenon which is stable in Standard German, but less frequent in the Swiss data, where genitive modifcation is partly replaced by prepositional phrases with *von*, as in the following example:

(7) Der_ART Bruder_NN *von_APPR* Frau_NN Dreifuss_NN ist_VAFIN übrigens_ADV Tierexperimentator_NN !_$.

The fact that the genitive case after prepositions is often replaced by a dative (e.g. *wegen, trotz*, see Table 4) in Swiss High German could lead one to expect higher counts of dative objects ($objd$), which is not the case. An important reason for the lack of increase is the fact that the dative object is itself under threat in German dialects, it is often replaced by the preposition *an*. German (like English) has a dative shift alternation, see e.g. Adler (2011). The difference is often seen as formal vs. informal, due to type of event or dialectal influence, also in German High German, but we find significantly higher counts for structures like the following in the Swiss corpus:

(8) Hier_ADV koennen_VVFIN Sie_PPER Ihre_PPOSAT Feriengruesse_NN *an_APPR die_ART Welt_NN senden_VVINF* ._$.

Although more research is needed on this question, such factors can explain the higher frequency of prepositions (tag *APPR*, dependency label pn). Coupled with the increased use of auxiliary verbs, Swiss High German also seems to be slightly more synthetic while German High German is slightly more inflectional.

Another observation related to language and dialect typology is illustrated in example (3), which shows a slightly unexpected verbal brace, where *zurückzuführen* does not appear at the end of the clause. While this choice can be caused by end-weight constraints, we wondered if there is a difference between Swiss High German and German High German. We counted how often objects and PPs of the main verb occur to the left or the right of the main verb. Frequencies to the right are considerably higher in the Swiss corpus, but closer inspection revealed that this is mainly due to the fact that there are fewer subordinate clauses in the Swiss data. In subordinate clauses, the default position of the main verb is clause-final, while in main clauses the position is after the subject or fronted elements (verb-second).

4 Conclusions

We have shown that automated data-driven methods manage to detect a wide range of the differences between Swiss High German and German High German described in the literature, by applying document classification and overuse metrics to a large web corpus, and shown that automatic part-of-speech tagging and syntactic dependency annotation detects patterns beyond lexis. Candidate lists from the relatively easy levels of lexis, but also from the more intricate levels of morphosyntax, syntax and style have been illustrated, also unveiling subtle stylistic differences. We

have pointed out strengths and possible pitfalls of the method. In future work, we plan to analyse the reported candidate lists in further detail.

References

Julia Adler. 2011. *Dative Alternations in German: the Argument Realization of Transfer Verbs*. Doctoral Thesis, The Hebrew University, Jerusalem.

Thomas M. Cover and Joy A. Thomas. 1991. *Elements of Information Theory*. Wiley Series in Telecommunications. John Wiley & Sons, New York, NY, USA.

Christa Dürscheid, Stephan Elspaß, and Arne Ziegler. 2015. Variantengrammatik des Standarddeutschen. Konzeption, methodische Fragen, Fallanalysen. In Alexandra N. Lenz and Manfred M. Glauninger, editors, *Standarddeutsch im 21. Jahrhundert – Theoretische und empirische Ansätze mit einem Fokus auf Österreich.*, Vienna University Press, pages 207–235.

Stefan Evert. 2009. Corpora and collocations. In A. Lüdeling and M. Kytö, editors, *Corpus Linguistics. An International Handbook, article 58*, Mouton de Gruyter, Berlin.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. Liblinear: A library for large linear classification. *Journal of Machine Learning Research* 9.

Kilian A. Foth. 2006. Eine umfassende Constraint-Dependenz-Grammatik des Deutschen. Manual, University of Hamburg: Fachbereich Informatik.

Dan Jurafsky and James H. Martin. 2009. *Speech and language processing : an introduction to natural language processing, computational linguistics, and speech recognition*. Pearson Prentice Hall, Upper Saddle River, N.J.

Marco Lui and Paul Cook. 2013. Classifying english documents by national dialect. In *Proceedings of Australasian Language Technology Workshop*. pages 5–15.

Christopher D. Manning and Hinrich Schütze. 2001. *Foundations of Statistical Natural Language Processing*. MIT Press.

Kurt Meyer. 1989. *Duden. Wie sagt man in der Schweiz? Wörterbuch der schweizerischen Besonderheiten*. IDS, Mannheim.

Bettina Rimensberger. 2014. Das Projekt Variantengrammatik das Standarddeutschen: Erste Forschungsergebnisse anhand deutschsprachiger Zeitungen. *Sprachspiegel* 4:102–110.

Christin Schätzle. 2013. *Eine computerlinguistische Untersuchung des Genitivschwundes*. Master's thesis, University of Konstanz.

Anne Schiller, Simone Teufel, and Christine Stöckert. 1995. Guidelines für das Tagging deutscher Textkorpora mit STTS. Technical report, Universität Stuttgart, Universität Tübingen.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*.

Rico Sennrich, Gerold Schneider, Martin Volk, and Martin Warin Warin. 2009. A new hybrid dependency parser for German. In C. Chiarcos, R. E. de Castilho, and M. Stede, editors, *Von der Form zur Bedeutung: Texte automatisch verarbeiten / From Form to Meaning: Processing Texts Automatically. Proceedings of the Biennial GSCL Conference 2009*. Tübingen, Germany, pages 115–124.

Claude E. Shannon. 1951. Prediction and entropy of printed English. *The Bell System Technical Journal* 30:50–64.

Elena Tognini-Bonelli. 2001. *Corpus Linguistics at Work*. John Benjams, Amsterdam.

Simone Ueberwasser. 2014. *Ein Requiem für den Genitiv? Vergleichende Betrachtungen zum Genitivgebrauch in den 1940er-Jahren und heute*. AkademikerVerlag, Saarbrücken, Germany.

Gertjan van Noord and Gosse Bouma. 2009. Parsed corpora for linguistics. In *Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics: Virtuous, Vicious or Vacuous?*. Association for Computational Linguistics, Athens, Greece, pages 33–39. http://www.aclweb.org/anthology/W09-0107.

Jeroen van Pottelberge. 2004. *Der am-Progressiv. Struktur und parallele Entwicklung in den kontinentalwestgermanischen Sprachen*. Tübinger Beiträge zur Linguistik 478. Narr, Tübingen.

Li Gong Yang, Jian Zhu, and Shi Ping Tang. 2013. Keywords extraction based on text classification. In *Advanced Information and Computer Technology in Engineering and Manufacturing, Environmental Engineering*. Trans Tech Publications, volume 765 of *Advanced Materials Research*, pages 1604–1609. https://doi.org/10.4028/www.scientific.net/AMR.765-767.1604.